Development of Geo-Knowledge Bases

Francisco J. Lopez-Pellicer

IAAA, Universidad de Zaragoza

July 8, 2009
GREASE-2

Introduction

Where I come from?

- **Advanced Information Systems Laboratory (IAAA)**
  - **http://iaaa.cps.unizar.es**
  - **R&D Group** Comp. Sci. and Sys. Eng. Dept., U. Zaragoza
  - **Research Lines** Spatial Data Infra., Loc. Based Serv., Geo. Info. Sys.
  - **Staff** (full time): 10 PhD, 14 master, 3 bachelor, 6 stagiaires
- **GeoSpatiumLab** (http://www.geoslab.com), Spin-Off, 14 Soft. Eng.

Research Stay at XLDB

- **Mid-March to mid-October**
- **Learning goal** $\Rightarrow$ **Understand the Geo-IR point of view** for my PhD thesis *Contributions to the problem of building a Geospatial Knowledge Base from deep Web sources*
- **Experience in** Enterprise-level **software development, geo ETL** processes and **ontologies**

Introduction

Work done

When	What	Specific context
March 2/2	Geo-Net-PT 02 (GKB 2.0)	
April 1/2	*Out of the town (Zaragoza)*	
April 2/2	WGO 2009 (GKB 3.0)	GikiCLEF 2009
May 1/2		
May 2/2		
June 1/2	Writting about GKB 3.0	XATA 2009
June 2/2	*Out of the town (GSDI 11th, Rotterdam)*	
July 1/2	Geo-Net-PT 02 (GKB 2.0)	

Geo-Net-PT 02

Context: a project unfinished

- **Geographic ontology of Portugal**
- Derived from **Geo-Net-PT 01 (Adm + Net)**, adds physical domain description
  - *Chaves, M.; Rodrigues, C. & Silva, M. J. Data Model for Geographic Ontologies Generation, XATA 2007*
- Status in March: partially loaded, several copies of the database, dump scripts in Perl, based on GKB 2.0

Initial goals

- Is all the **physical geo data loaded**?
- Load **administrative geo data**
- Compute **spatial relationships**
- Create **dumps** in OWL format using the scripts

Geo-Net-PT 02

New issues

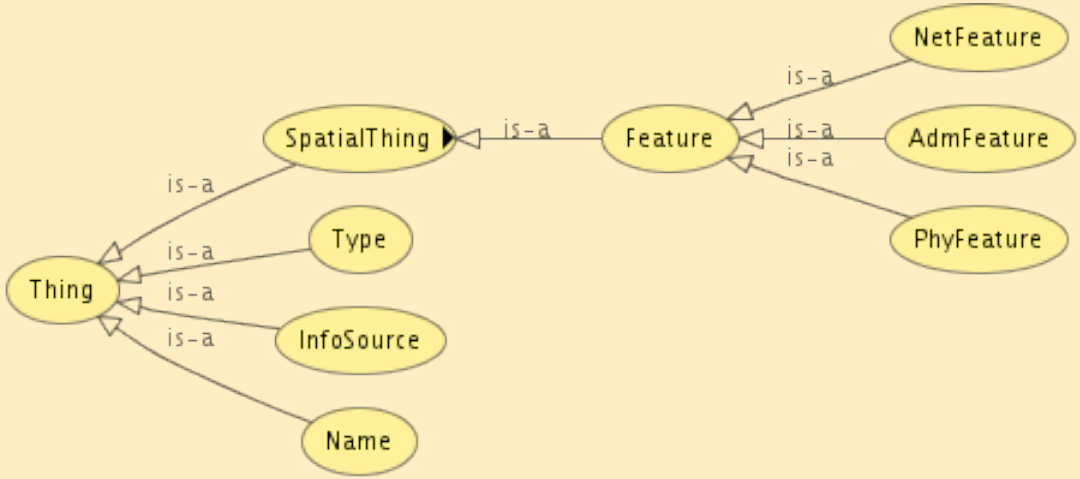
- Does the legacy database support for multiple **projections**: 5 2D + WGS 84 (Geodetic)?
- Does the available **Perl scripts** generate a complete dumps of the model?
- How to create **maintainable code** (maestrado level)?

Solution

- Add **support** for store geometries of different **projections** (in ADM only)
- Create a set of **Java tools**: Load geometries, compute relationships, new GoG
- Use **Java** code
  - API generated using an ORM library with spatial extensions (Hibernate Spatial)
  - Jena + TDB as data backend for GoG (no database required!)

Geo-Net-PT 02

TBox



ABox

```
gn:AF_1
    rdf:type gnt:AdmFeature ;
    dc:title "Abrantes"@pt ;
    gnt:hasPreferredName
        gn:AN_1 ;
    gnt:hasSource gn:IS_100 ;
    gnt:hasType gn:AT_CON ;
    gnt:population 42235 .
```

Geo-Net-PT 02

Game over?

- Not yet, this is the next step in this work
  - Relevant **feature types missing** or incomplete
  - **Names** should be cleaned and pruned
  - **Taxonomy** should be also revised
  - **OWL dump** should be verified

WGO 2009

Context: GikiCLEF 2009

- **World Geographic Ontology**
- GikiCLEF 2009 is an evaluation task under the scope of CLEF.
- Its aim is to evaluate systems which find **Wikipedia entries / documents** that answer a **particular information need**, which requires **geographical reasoning** of some sort

Special requirements

- **Multilingual** task
- **World** wide coverage
- Answer are **Wikipedia** documents

WGO 2009

Alternatives ...

1. Reuse WGO/Geo-Net-PT
2. Create a (geo-)ontology from categories
   - *Suchanek, F. M.; Kasneci, G. & Weikum, G. Yago: a core of semantic knowledge, WWW '07*
3. Create a (geo-)ontology from infoboxes and templates
   - *Wu, F. & Weld, D. S. Automatically refining the wikipedia infobox ontology, WWW '08*
4. Create a (geo-)ontology from text
   - *Nguyen, D. P. T.; Matsuo, Y. & Ishizuka, M. Exploiting Syntactic and Semantic Information for Relation Extraction from Wikipedia, TextLink'07*

WGO 2009

... for a 1 man-month task

- Only from **SQL** dumps
- Development in **parallel** with the users
- It is **unclear** which information is relevant

Extracted data

1. Identify from **templates** pages that are about features
   - 82769 features which contains coordinates (Success!)
2. Identify from these pages its **categories** and build a taxonomy
   - 131882 features but a noisy taxonomy derived from +2000 categories (Failure!)
3. Add additional info from **page links** (other languages, related pages)
   - 210401 names (Success!)

WGO 2009

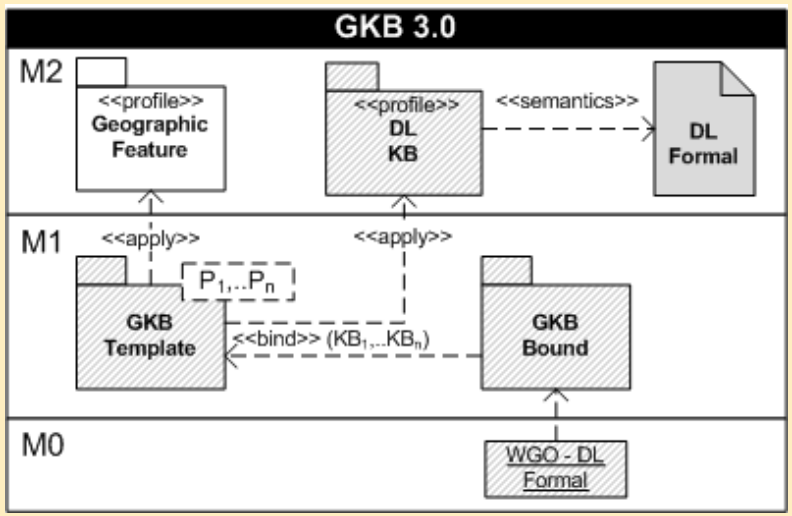
Some thoughts about part-of (or other) relationships

- Is possible to derive Part-of relationships from wikipedia categories
  - ".../Category:Municipalities_of_Portugal" can be rewritten as $Municipality\,(x) \wedge IsPartOf\,(x, Portugal)$
  - and this page has the subcategory of ".../Category:Municipalities_of_the_Algarve" that can be rewritten as $Municipality\,(x) \wedge IsPartOf\,(x, Algavre) \wedge IsPartOf\,(x, Portugal)$ and adds the info $IsPartOf\,(Algavre, Portugal)$
- However there are problems
  - ".../Category:Municipalities_of_Portugal" contains the subcategory of ".../Category:Lisbon" that must be rewritten as $Related\,(x, Lisbon)$ (but adds the info $IsPartOf\,(Lisbon, Portugal)$)
  - ".../Category:Regions_of_Portugal" contains the page ".../Administrative_divisions_of_Portugal" but the best most suitable assertion is $\neg\,(Region\,(x) \vee IsPartOf\,(x, Portugal)) \wedge Related\,(x, Portugal)$

WGO 2009

Re-engineer the storage

- GKB 3.0 $\Rightarrow$
  - Model all the relationships as **M:N**
  - Each concept is **linked to data in the WWW**
  - Feature types are sets of things and its relationships might imply set containment
- GKB 3.0 API $\Rightarrow$ **fast basic transitive inference** for **is-a** and **part-of**
- Without breaking other people work!



Future work

Ontologies, wikipedia, GKB

- Ontologies
  - Can the generated **ontology schema be improved**?
  - Which parts of the model should be stored in the ontology as **annotations**?
  - Which **explicit relations** among features should be supported? How to enforce them?
  - **Feature types** as instances or as classes? Both? Configurable?
  - Why OWL? Why not **SKOS**?
- Wikipedia
  - It is possible to build a **geo-ontology** from the Wikipedia?
- GKB
  - Can GKB 3.0 **replace** GKB 2.0?
  - Relational **storage**?
  - GKB **API** queries, **SQL** queries or **SPARQL** queries?